BACTERIOLOGICAL REVIEWS, Dec., 1965 Vol. 29, No. 4

Printed in U.S.A.

Symposium on Information Science

I. International Aspects of Information in Microbiology[1]

P. R. BRYGOO

Institut Pasteur, Paris, France

INTRODUCTION 506
EVOLUTION OF BIOLOGICAL SCIENCE INFORMATION 507
MICROBIOLOGICAL SCIENCE INFORMATION TODAY 509
PROPOSED COORDINATION OF INFORMATION SERVICES 511
Primary Level of Service 511
Secondary Level of Service 511
Tertiary Level of Service 511
MACHINE RECOGNITION AND RETRIEVAL 512
LANGUAGE DIFFICULTIES 514
LANGUAGES OF THE FUTURE 515
LITERATURE CITED 515

INTRODUCTION

The International Association of Microbiological Societies (IAMS) in 1958 created a Permanent Committee for Microbiological and Immunological Documentation, with the intention of establishing an international pilot center for research in information methods for microbiology (1, 2). This IAMS section has pointed the way since its inception. It has managed, practically without financial means, to establish connection with 34 national microbiological societies. As the first modest fruit of its labor, it will soon be able to present a list of the primary journals considered by microbiologists to be the main sources of information in microbiology. This list is being compiled in Sweden by the Committee's Secretary-General, G. Tunevall, in collaboration with correspondents designated by the 34 national microbiological societies. The list will give not merely the essential details for all primary journals that print articles of interest to microbiologists, but will also indicate the secondary services that cover them. This list represents the first step in a long-term program the purpose of which is to interconnect the abstracting services and to assure the proper indexing of their content for automatic treatment by computer.

For these projects, no financing mechanism has as yet been found. This fact should not make us abandon the task or discourage us about the future. At present, on all sides, attention is being drawn to the importance of working out new means of communication. The new awareness of the need for these means must be followed by careful work in information research. This awareness is no doubt a sign that the moment is close when the absolute necessity of international effort in scientific communication within specialized fields will be understood.

The present symposium demonstrates the alertness of the American Society for Microbiology (ASM) to information problems and a growing desire for solutions. This symposium will, we hope, be a precursor to similar efforts by other national microbiological societies. The reception accorded by the ASM to this exposition of general aspects of information for microbiology is encouraging. Microbiology is one of the most dynamic disciplines in biology. It will be a leader and an example in that it will contribute thoughtful information research and seek means to solve its information problems. Only thus can it assure its future in the progress of science.

We are concerned with the ways in which scientists communicate with each other. This complex problem has called up for reconsideration so many habits that the change is better described, not as an evolution, but as a revolution. As scientific communication problems change, the tried methods and established concepts applied to them continue to be adjusted empirically. They attain that degree of perfection possible within the limitations of the technical development of the time. Meanwhile, new approaches to entirely unexplored realms are opening to us. We should consider these as a chal-

[1] A contribution to the symposium "Information Science" held at the Annual Meeting of the American Society for Microbiology, Washington, D.C., 4 May 1964, with H. W. Batchelor as convener, and G. H. Nelson, M. D. Nelson, and A. J. Shanahan as consultant editors.

lenge to us as scientists. We should enter them with methodical boldness and we should not be discouraged by the effort we know will be required of us.

What determines the power and the influence of communities is the dynamism of their scientific and technical life. In consequence, the communication and circulation of scientific information have taken on considerable practical importance, in some cases greater than that of the original generation of this information. Thus, on a universal scale, a new and absolute necessity has arisen, i.e., the obligation to participate in the exchange of knowledge. Within this framework each partner strives either to keep or to acquire an advantage and to enrich, more and more rapidly, a common fund of perpetually expanding knowledge which is the world's most inexhaustible natural resource.

Under these circumstances, it is not surprising that the quantity of communications of scientific interest should attain such a growth rate. This rate long since has overwhelmed the capacity of even specialized services (which were created only a few decades ago) to solve the information problem for a limited community. Almost every scientific discipline has numerous services specialized in information treatment. These services, even those with considerable financial means, find it more and more difficult to accomplish their task. Just as, at an earlier date, the problem of scientific information came to exceed the capacity of individuals, so too, at present, the problem tends to overwhelm the resources of organizations at the level of national scientific communities.

This crisis has incited (besides impressive numbers of attempts merely to describe it in statistical estimates) a widespread hunt for new solutions (based on entirely different principles) that could assure the collection, processing, organization, selective search, and distribution of scientific information. It would seem reasonable to hope that treatment of scientific information, at first by use of punched cards (whether by hand or mechanical methods) and later by use of the much more powerful resources of electronic computation, could reach a level of industrialization and automation allowing a speed and a yield unattainable by manual and human intellectual means alone. My role here is not to recapitulate the thousands of research efforts that have been made during the last 10 years with a view to mechanizing and automatizing documentation. I shall only remark that, although they have not yet yielded results offering a complete and general solution to the documentation problem, they have contributed to simplifying the execution of numerous routines and, at least within certain limits, have made possible by automatic means work not feasible by hand methods.

The important point to note is that, although this line of research has not yet yielded all the fruits one can and should expect from it, a revolution is under way. Nothing suggests that it is going to stop. In fact, automatic data processing should reach sooner or later those most sophisticated accomplishments: automatic reading, automatic translation, automatic abstracting, automatic indexing, automatic classification, and perhaps one day even automatic reasoning, automatic learning, and application of strategy games to the management of research. Needless to say, these accomplishments can be attained only through the research teamwork of scientists and cannot be expected by a scientific community passively awaiting them as a consequence of the evolution of "hardware."

Science information specialists in the US, in the USSR, and in Europe are becoming more and more aware that their future and that of scientific progress are closely linked with the development of applied mathematics, of symbolic logic, of linguistics, and of computer programming. Moreover, the results of the effort that these specialists and generalists have devoted show clearly that they count on research oriented in this direction.

Scientific activity has assumed a new character, marked essentially by the universality of its object and its method and by the brief duration of useful life of the data it permits us to acquire. The communication network which should keep such scientific activity coherent must itself be adapted to these exigencies and thus keep pace with the most advanced treatment of the raw material of information.

Evolution of Biological Science Information

What, precisely, is happening in scientific information in biology, and in some subareas that are crossroads of major importance, of which microbiology is a particularly interesting and demonstrative example?

Fifty years ago, microbiology occupied scientists in a handful of large research institutes in a few large countries. Their work was published in a limited number of journals in French, English, or German. A very few documentation services throughout the world easily covered all the work they published and gave excellent abstracts of it in each of these languages. These abstracts were generally read with interest and attention by all microbiologists who thus, at an acceptable input of effort, had a general and synthetic view of research progress in their field. Microbiology was essentially a science of medical and veterinary application oriented toward infectious diseases and their prophylaxis, as was natural in view of the nature and urgency of the problems at that time.

In 1930, only 34 years ago, the first International Congress for Microbiology was held in Paris. For us, now, it seems hard to imagine that almost all the participants at that meeting knew each other personally, by their work and their correspondence.

After World War II, the situation was drastically changed: international communications had been cut; a new generation of research workers had replaced the former ones; a huge number of laboratories and research organizations had been established; and new specialties had been born of which war-torn Europe had barely heard. The former harmony in knowledge and its communication was replaced by a marked unbalance. In addition, the fields to be covered had proliferated to such an extent that the size of the research community required to cope with the proliferation could not possibly be forced into the existing communications framework.

The European information services such as *Zentralblatt* and the *Bulletin de l'Institut Pasteur* had been left high and dry in this evolution. However, other work tools (especially in English, which had become the prime international scientific language) came to the rescue of the information sciences with first-rate activities. This was the case, in particular, for *Biological Abstracts* and *Excerpta Medica*, more recently created. The different abstracting services published by the Commonwealth Agricultural Bureau made an honorable effort to keep their standing. Meanwhile, ambitious and recent national organizations, such as the All-Union Institute of Scientific and Technical Information (Vsesoyuznyy Institut Nauchnoy i Tekhnicheskoy Informatsii or VINITI), in the USSR, and the Centre National de la Recherche Scientifique (CNRS), in France, began to develop large scientific information services of encyclopedic intent: the *Referativnii* in the USSR and the *Bulletin signalétique du CNRS* in France.

Without a doubt, each of these documentation services has a reason to exist, corresponds to needs that are not identical, and defends traditions worthy of high esteem. Notwithstanding this, they are of unequal value and, moreover, considerable gaps exist among them. None of them, particularly in microbiology, can pretend to give a complete and faithful documentation of all the work meriting interest.

Their irreducible particularity is often criticized, but it arises from numerous technical causes that are inherent in the organization that produces each instrument. In consequence, all attempts at genuine international coordination directly between services have been fruitless, in regard to making their content and format really exchangeable.

A few facts and figures will help to depict concretely the problem we face at present. Figure 1 depicts the effort made (between 1950 and 1960), by four large abstracting services, to keep up with the explosive growth of original scientific articles. During these 10 years, the number of annotated bibliographic references presented annually by *Chemical Abstracts*, the *Bulletin Signalétique du CNRS* (parts 1 and 2), and *Biological Abstracts* has exhibited an exponential growth rate. Thus, this growth is not a figure of rhetoric, but a fact. In regard to the biological sciences, we note that the *Bulletin signalétique* (in biology) and *Biological Abstracts* (starting with 30,000 to 35,000 references yearly in 1950) furnished about 85,000 to 90,000 abstracts yearly in 1961, i.e., they nearly tripled their activity. Since 1961 the figures have continued to rise, to around 100,000 abstracts in 1963 for *Biological*

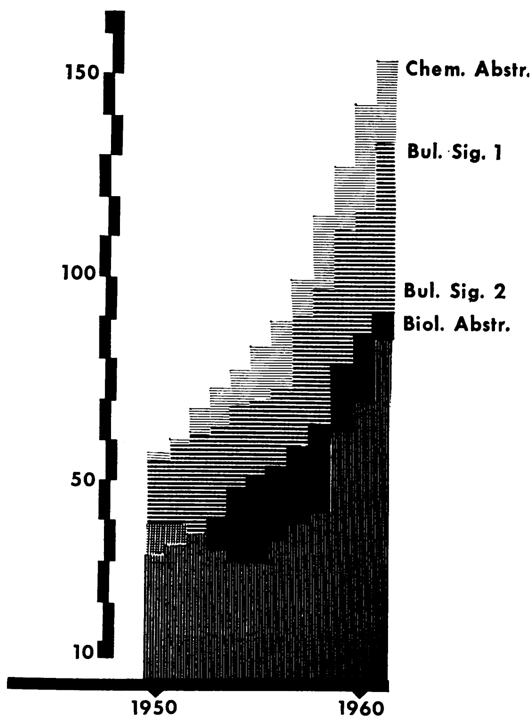

FIG. 1. *Thousands of annotated bibliographic references, Chemical Abstracts, Bulletin Signalétique du CNRS, and Biological Abstracts.*

Abstracts and to nearly 120,000 abstracts in the same year for the biology section of the *Bulletin signalétique*. The volume of the latter has thus quadrupled in 13 years and increased 33% during the 2-year period, 1962–1963. According to the figures given for the US by the National Federation of Science Abstracting and Indexing Services, the total number of abstracts and citations provided by members of this Federation for 1963 was around 950,000 items, or more than double the number of items in 1957. This figure (which unfortunately does not take overlapping into account) gives an idea of the global quantitative aspect of the scientific information problem. It is comparable approximately to the figure given by the VINITI in the USSR, which, it seems, managed to cover about one million scientific and technical articles in 1963.

Microbiological Science Information Today

As regards microbiology, immunology, and related fields, it is difficult to give generally valid figures, but it seems probable, according to my estimate, that the fields concerning microbiology account for at least 15,000 to 20,000 documents of those that are abstracted yearly, i.e., about 15% of the number published for biology as a whole. These figures agree rather closely with estimates made in the US by *Biological Abstracts*. It is to be noted that almost no documentation service specialized in our field covers any such number of articles. The one coming closest is the *Bulletin de l'Institut Pasteur*, which publishes over 14,000 references annually. We must therefore conclude that, if the general abstracting services (like *Biological Abstracts* and *Bulletin signalétique du CNRS*) are already very insufficiently complete, the specialized services are even more so.

A very instructive example in our domain was

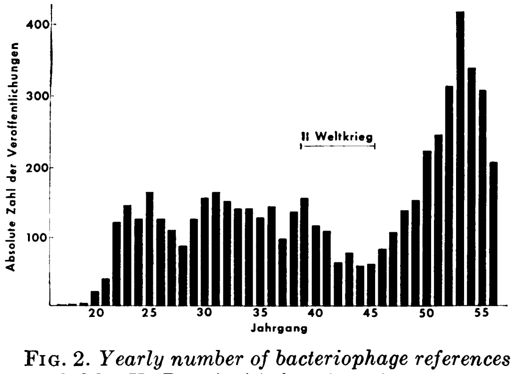

FIG. 2. *Yearly number of bacteriophage references compiled by H. Raettig (7) for 1917–1956.*

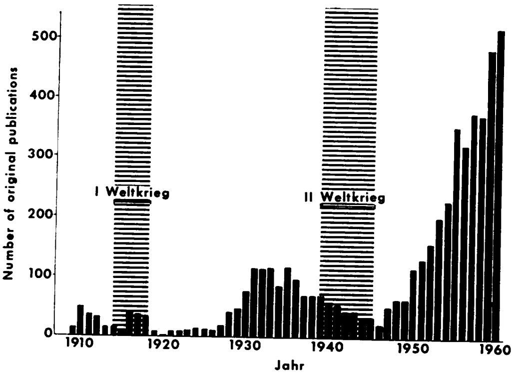

FIG. 3. *Yearly number of poliomyelitis immunity references compiled by H. Raettig (8) for 1908–1961.*

furnished by H. Raettig in Germany in connection with two bibliographies that he compiled, one for bacteriophage (7) and the other for immunology in poliomyelitis (8). For these two subjects, Raettig used a bibliographic research technique similar to the one worked out by Eugene Garfield for citation indexes (6). The method made possible retrospective constitution of an extremely large body of original references, directly from original articles (Fig. 2). The periods covered were, for bacteriophage, 1917 to 1956 (40 years) and, for immunology in poliomyelitis, 1908 to 1961 (52 years). The last three years do not count, since the method used does not permit obtaining significant figures in this area.

The prodigious increase in the number of original articles on phage between 1945 and 1953 is noteworthy. The annual volume of literature increased about sixfold in 8 years after the second World War. This phenomenon corresponds to the expansion in basic research in virology.

Concerning the original articles on immunology in poliomyelitis, a more striking, but similar, phenomenon is noted between 1945 and 1960 (Fig. 3). The annual volume of original publications increased more than 12-fold in 15 years after the second World War. This increase corresponds to the growth of both basic and applied research in virology in a particularly active sector.

Although the figures for 1960 probably are less than the real figure (for the technical reason noted above), it could be objected that the picture as given in these two examples is false: firstly, the domains selected are among those most in research fashion during the last 10 years; secondly, the reference year 1945, just after the war, represents an abnormal point of depression consequent to exceptional circumstances. These objections, although meriting discussion, in fact do not have much bearing on the points we will discuss later. Preferentially, we will examine further the second of Dr. Raettig's examples, i.e., immunology in poliomyelitis (Fig. 3).

On the basis of the original articles forming this body of bibliographic references, a year-by-year inquiry was made to establish the percentage of these references that had been covered by different abstracting services existing at the time and used in Germany (Fig. 4). This study concerns especially the main specialized German abstracting journal, *Zentralblatt für Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene* (Abteilung I, *Referate*), which was followed from its inception to the present. Figure 4 also gives information on two other German services (*Zentralblatt für die gesamte Hygiene und ihre Grenzgebiete* and *Kongresszentralblatt für die gesamte innere Medizin*) that ceased publication after the second World War and have little interest for us. Finally, results are shown, in the period after World War II, for two English-language services: *The Bulletin of Hygiene* and *Excerpta Medica*, Section IV.

Figure 5 shows the coverage in detail. The body of references collected from original articles arbitrarily is taken to represent 100%. We see, separately, the percentages of these references found in *Zentralblatt*, *Bulletin of Hygiene*, and *Excerpta Medica.*

In the case of *Zentralblatt*, three periods are characteristic of the evolution of the problem. Before World War I, the service was about 70% efficient. After World War I, more than 10 years of effort was needed to attain approximately 40% efficiency (in 1930). Between 1930 and 1940, the average reached approximately 50%. After World War II, *Zentralblatt* coverage was less than 20%. Despite the efforts made to catch up, the journal cannot be considered to be an instrument

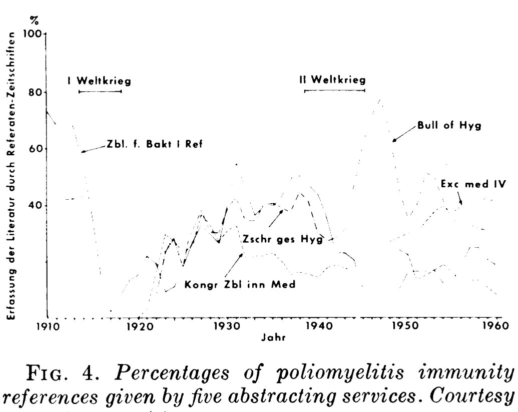

FIG. 4. *Percentages of poliomyelitis immunity references given by five abstracting services. Courtesy of H. Raettig (8).*

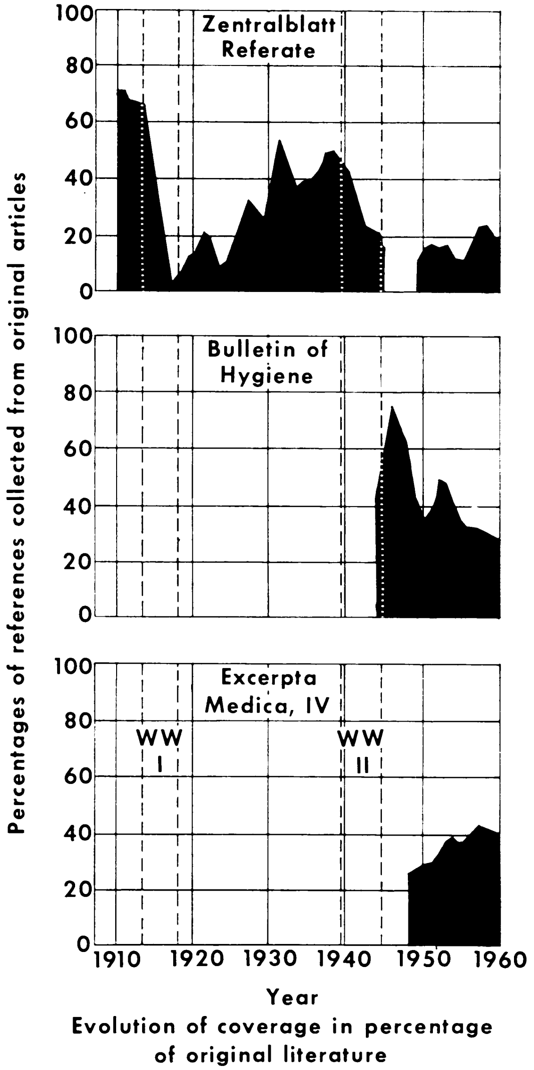

FIG. 5. *Percentages of poliomyelitis immunity references cited by three abstracting journals. Number of original articles taken as 100%. (From data of H. Raettig.)*

of bibliographic research, as 80% of the published literature escapes its coverage.

It is also interesting to note that the *Bulletin of Hygiene*, whose coverage was very satisfactory just after the second World War, declined thereafter, and that, during the last year shown, its coverage did not attain 30%.

Excerpta Medica, more recently created, progressed for a certain length of time and reached approximately 40% efficiency.

The example presented shows that the development of the original scientific literature has not been followed since World War II by a proportional development of specialized abstracting services. It is simpler still to show that the increase in coverage for most of the specialized services in microbiology has been feeble in comparison with that of the general services. The specialized abstracting journals have continued to appear in print and grow, but not in proportion to the growth in the number of original articles. Thus, their adequacy has deteriorated.

It seems of interest to compare the curve for the number of original articles on the immunology of poliomyelitis during half a century (Fig. 3) with a figure representing the successive creation of different scientific information services of at least partial interest to microbiology (Fig. 6).

Development of abstracting services marked the first half of the 20th century. Their progression has lead to a paradoxical and somewhat disconcerting situation: whereas the value of each of them tends to deteriorate, their number tends to increase. Theoretically, the investment of means is made for the purpose of recording and grouping information into wholes (for a particular specialty's needs) that should be as complete and as coherent as possible. But, in fact, the investment is absorbed by the development of services that are more incomplete and more and more dispersed.

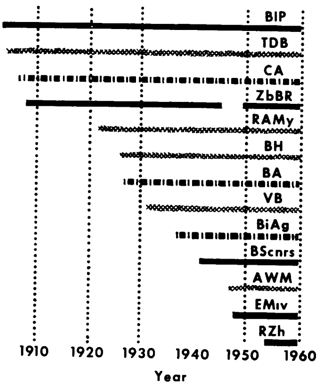

FIG. 6. *Establishment of information services. BIP, Bulletin de l'Institut Pasteur; TDB, Tropical Disease Bulletin; CA, Chemical Abstracts; ZbBR, Zentralblatt für Bakteriologie (Referate); RAMy, Revue of Applied Mycology; BH, Bulletin of Hygiene; BA, Biological Abstracts; VB, Veterinary Bulletin; BiAg, Bibliography of Agriculture; BScnrs, Bulletin Signalétique du CNRS; AWM, Abstracts of World Medicine; EMIV, Excerpta Medica (IV); RZh, Referativnii Journal.*

PROPOSED COORDINATION OF INFORMATION SERVICES

Primary Level of Service

We regard the publication of original documents as the primary level of treatment or service.

Secondary Level of Service

Before World War I, a maximum of four secondary information services were published (*Bulletin de l'Institut Pasteur*, *Tropical Diseases Bulletin*, *Chemical Abstracts*, and *Zentralblatt*), each of which covered easily the scientific literature concerning its specific discipline and, in fact, covered the needs of microbiologists as a whole. At present, one could easily name at least 15 different services that do not coordinate with each other and that do not satisfactorily accomplish the task that has become overwhelming in respect to their individual means.

This situation raises the need for coordination of abstracting services in order to constitute pools of capsule information that would remedy the present dispersion and incoherence of bibliographic services.

The contents of existent abstracting journals may be looked on as partial, intermediate memories which it would be desirable to explore simultaneously, but with regard only to that unique fraction which each alone possesses. The coverage of these abstracting journals overlaps heavily and is incomplete. On the other hand, each service makes its contribution of information which does not appear elsewhere, in undetermined proportions, but which justifies *a priori* that interest be accorded to each of them. We must admit the dispersion of resources invested in multiple abstracting services. It would be vain for us to try to reorder these resources; therefore, we must use this fact as a basis of information-science research designed to make coherent what is not coherent as matters now stand.

Tertiary Level of Service

We must attempt to link the annotated-bibliographic information services (commonly known as "secondary"), or abstracting services, at a third level that I propose to call the "tertiary level" of treatment of scientific information. This tertiary level, as we can easily imagine, is by nature international and assumes the existence of a single international organization for information in a given specialty. Assuming the usefulness of such a tertiary service to subspecialties like ours, we shall consider some of its technical aspects, in order to identify difficulties to be overcome. Three functions of this tertiary-level organization would be to:

(i) Reduce to a single whole the content of the partial memories of many information services. Each original (primary level) document covered by at least one of the abstracting services (secondary level) would thus be entered in the collective memory at the tertiary level. Thus, during subsequent searching of the content, no original article covered by any service could be lost. On the other hand, original articles covered by two or more services would be entered only once in the tertiary memory, although the identifying coordinates ("addresses") of all documents in the secondary memories would be stored. This procedure would make it possible to have only one "address" as a key to several others and to reduce to minimal volume the content of the tertiary memory by eliminating all redundancy and all overlapping of the secondary memories.

(ii) Eliminate from this single series that fraction of the partial memories recognized as nonrelevant. This function would eliminate, at the origin, the fraction of the documents covered at the secondary level that does not concern the specific discipline for which the tertiary level information pool is established. The obvious purpose is to limit the tertiary memory to useful content only, i.e., usable by the specialty it serves. For example, a great many original articles covered by *Biological Abstracts* do not concern microbiologists. *A priori*, such a preselection of content is necessary whenever the secondary service is not specialized in strictly the same domain as the intended users of the tertiary memory.

(iii) Index the documents covered in the tertiary memory in a consistent manner. By adopting a "documentary language," a single intermediary language can be adapted to the particular needs of the discipline that the tertiary organization is to serve. In our case, the disciplines are microbiology and its relatives. This condition is absolutely essential to obtaining from the system an acceptable yield when the content is searched in regard to a specific question.

The first two functions concern information storage, whereas the third function concerns information retrieval (recall). What practical approaches to the problems posed by the first two functions are conceivable? As regards selection we shall discuss two approaches, one based on subject content of the abstracts and citations, the other taking as the starting point a list of journals established by the specialty and thus considered to contain the articles relevant to it. Each of these has its advantages and disadvantages.

The surest method is doubtless to select the material for relevancy by directly screening, item-by-item, the content of the secondary memories. This procedure would mean screening the entire content of the secondary memories or, when the memory is already organized in preselected subject sections, screening part of the secondary memory. The validity of this preselection, from the specialty's viewpoint, would need to be determined.

Another, somewhat different, approach could be based on a pre-established list of primary journals recognized by the specialty to be of interest for it. All original articles from these primary journals, but only these, are selected from the secondary service "memory" whenever a citation from these journals appears. The disadvantage of this method is that it, too, requires screening the whole content of the secondary memories. Its advantage is theoretically to permit automation. The material so selected from a list of journals must nonetheless still be screened item-by-item for relevance to the field to be covered. There is merely less of it to screen. However the selection operation for input into the tertiary memory is performed, it requires the constant work of specialists trained to distinguish, in each secondary memory, what belongs and what does not belong to the domain served, e.g., microbiology.

The next stage is storage in the tertiary memory. Here arises the difficult problem of identifying primary articles that two or more secondary services have recorded. In practice, this identification problem has been solved at the tertiary information level in certain highly specific fields, for example, by the Food and Agricultural Organization of the United Nations (FAO) in Rome, Italy, for documents concerning fisheries. In the FAO system the volume of documents is not very large. Large volumes of information cannot feasibly be treated in the same way. If we pool the content of several heavily overlapping secondary memories, each possibly containing over 10,000 to 15,000 references per year, only machine methods reasonably can be expected to identify multiple citations as pertaining to a single original article. On the scale applicable to documentation in microbiology, the essential preliminary to any attempt to establish a collective memory is to work out a procedure for identifying references by machine. Let us, therefore, examine this problem carefully.

Machine Recognition and Retrieval

Suppose that we have two citations of the same original article, one from *Biological Abstracts*, the other from the *Bulletin de l'Institut Pasteur*. We want a machine to recognize that they refer to the same article, although the form of the citations differs. The machine either must receive an identical symbolical representation of both citations or else be able itself to reduce to identity two different representations. This basic point may seem simple. In practice, it raises considerable difficulty. Only appropriate research can show whether the difficulty can be surmounted in all cases. At least one thing is certain: references in the form given by secondary services can be introduced into a machine but, because the standard form for each service is different, one set cannot be matched by machine with another set of different origin. One must therefore find a symbolic expression that can be extracted or derived, with no error, from all the forms that bibliographic references referring to a single document may take in different secondary information-service citations. At the same time, this expression must remain specific enough so that the probability is almost negligible that, from secondary service references referring to different original documents, it be obtained purely by chance.

Only when such a universal machine code for references has been defined will it become possible to search out automatically the overlapping in content of different secondary memories in order to eliminate duplications in a tertiary memory. At the start, the address of the reference in the secondary memory and the universal symbol derived from this reference will have to be coded manually. However, some secondary services already produce their references in machinable form, on punched cards or punched tape; in this case, the universal symbol could be obtained automatically by a specially programmed computer. Preliminary research for computer programming must be done with such machinable reference material. IAMS Permanent Committee for Documentation in Microbiology and Immunology suggested that this research might be undertaken cooperatively by *Biological Abstracts* and the *Bulletin de l'Institut Pasteur* with the machinable references they produced. Obviously, if this research project yields the hoped-for result, its practical implications would be considerable; reliable machine techniques for pooling references would be applicable in general and would be put to use, especially by many subspecialties. It would be a big step forward if means were available to assemble this single series of symbols that represent the primary documents and of linking these symbols by an address to the more complete representation given in citations and abstracts.

Next, a coherent retrieval system must be found. This step is by far the most difficult. It raises the problem of indexing the whole content of the system by means of a single documentary language. As Fairthorne (5) put it: "Indexing is the basic problem as well as the costliest bottleneck of information retrieval."

As concerns annotated-bibliographic documentation, information retrieval encompasses those operations required to find and assemble the bibliographic references of documents dealing with a definite subject at a particular time. Indexing is the function that permits one to make this selection without having to examine every document in the collection. Whatever the method or technique of indexing, the principle is always the same: to establish classes in which documents fulfilling certain criteria can be found, so that, when a search involving these criteria is made, no class other than the relevant one need be searched. A bibliographic reference introduced into the collection will be found only if it has been included in one of the classes searched, i.e., only if, at the time of indexing, it was attributed to this class. Taken together, the index classes, their distribution, and the formal relationships linking them together constitute what is called the indexing language or documentary language of the system. The concepts that define index classes constitute the vocabulary of this language; the various relationships of classes to each other are its syntax.

Whatever document is to enter the system must be represented according to the rule of this language. All subsequent operations (searches in particular) will be performed by the machine, not on the document itself but on the symbolic representation given it by the indexer within the norms of the documentary language used.

To enable search for documents, the "question" asked must also be transliterated to the machine, i.e., be represented in conformity with the symbols of the documentary language. The "question" put to a machine system is none other than the "ideal" description of the documents wanted, which may be in the collection and which, when they entered it, received a corresponding description.

Between the unknown documents which it is useful to find, and the genuine and complete intention of the question posed, there is only a degree of analogy, but never complete coincidence, except when all the characteristics of a desired document are perfectly known in advance and one wants that document in particular and no other (a case not representing a true bibliographic search).

To the extent that the selection operation yields a list of documents, as a "reply" to a "question," it is an operation of rigorous logic, dependent on the ability to write an equal sign between symbolic expression for content of the question and symbolic expression for content of the reply. Very naturally, the actual content of the document and the actual intention of the question are quite different matters from their symbolic expression; between the content and the intention themselves and the symbolic expression that transliterates them, there is room for all sorts of choices and approximations. Here enters the notion of degree of analogy.

LANGUAGE DIFFICULTIES

In the measure that operations performed on documentary languages constitute operations of symbolic logic, like all forms of mathematics they can be described as the art of conducting in a rigorous manner a series of exact operations on data that, themselves, are only approximations; we pass from a natural nonrigorous language to an artificial rigorous conventional language designed to serve as data for computation. Thus, we see that there is no obligatory relation between methods and techniques used for setting symbolic expressions, and methods, techniques, and instruments applied in computation on these symbols. In fact, if we set aside for the minute all idea of convenience, speed, and cost, symbolic logic can just as well be applied in operations with manuscript file-cards in pasteboard boxes as with magnetic tapes; additions can be done just as well on a slate or an abacus as on an electronic computer. This is why we say that the instrument used in manipulating symbols is of secondary importance.

What *is* a primordial necessity in passing from a natural to a documentary language (as in expressing the value of physical data by a measure and the choice of a unit) is perfect agreement on the conditions of approximation in which one measures the values to be treated and on what system of symbols is going to be used to express them. There is a perfect logical identity between symbolic representation of a question and symbolic description of a document when, on comparison of the two, the document is selected in reply to the question. This means, not that perfect concordance exists between the document's actual content and the question's real intent, but only that the degree of analogy between the two is sufficient. To understand this point better, let us suppose we are searching for physical objects of a certain length by means of a documentary language that is merely the numerical measure of this length. Suppose that objects whose lengths "correspond" to that of a standard object are to be selected. According to the exactness of the measurements we have made of (i) the objects in our collection, (ii) the standard object, and (iii) the precision of the numerical representations we have assigned to them, the degree of correspondence will vary considerably, and the number of objects included in the class "corresponding" will vary likewise.

Probably some references furnished by the machine in answer to a question will be "noise," i.e., nonrelevant, because their actual content will not be directly related to the actual sense of the question. When the symbolic representation of a question and that of a document are different, i.e., when, on comparison of the two symbolic expressions, the document is not selected, this nonselection does not mean necessarily that the document's content did not agree with the intent of the question. It means only that the difference was sufficient, in view of the characteristics of the conventional language used, for document and question each to have different symbolic expressions. Thus, some documents will not be recalled from the memory despite the relevance of their actual content to the actual sense of the question. The notion of richness of reply, or recall ratio, corresponds to the width of wave range of the system, as opposed to the purity of reply, or relevance ratio, which corresponds to the system's selectivity.

For any given system, richness of reply varies inversely to purity of reply, recall ratio inversely to relevance ratio. This means that the more selective we make a documentary language, the less will be the "noise" or nonrelevant material selected, but the price will be nonselection of a large number of useful documents. On the other hand, an exhaustive language will yield most of the useful documents, but at the price of much "noise." Some information scientists, I among them, had felt that these principles were probably correct and of considerable theoretical value. We were glad indeed to see them proved experimentally by the research of C. W. Cleverdon (3, 4) and the Cranfield group in their work on index languages.

To have a rich memory at our disposal is obviously a prime requisite for success in documentary research, but it is not at all a guarantee. The value of recall obtained from a memory of a given content depends on how we organize as well as on how we explore the memory. Quite naturally, we might suppose that the decisive advantage of one system over another lies in the choice made between hierarchical, alphabetical, or facet classification or coordinate indexing; this assumption is, however, false. Having the most powerful computers will not suffice, either, to solve our problem. The prime necessity is to possess a documentary, or indexing, language perfectly suited to our own need and use. Above all, it must be consistent, i.e., admit only a single symbolic vocabulary system and a single syntax. This language, or indexing system, must be applicable, with a change in machine search programs, to the different levels of generality and specificity required. As Cleverdon and Mills (4) have proved, only very deep indexing of documents in a very specific indexing language can make this possible.

LANGUAGES OF THE FUTURE

To coordinate existing abstracting services and to pool their information is unimaginable. We can imagine a tertiary memory containing the sum of secondary level information, but, to explore it, the only possibility is to reindex its content by means of an appropriate documentary language. Because this language must be rigorously exact, it cannot be a natural language. The starting point in constructing it can be words from a natural language, e.g., English, but like all documentary, or indexing, languages it must be conventional as are nomenclatures in biology and chemistry. It so happens that languages of this type are the ones best suited both to international exchanges and to mechanical handling. For international technical and scientific communication, they are the languages of the future.

In this connection, an immense amount of work awaits doing in microbiology. The common fund of scientific information can be treated for our use; abstracting service material can be taken simply as raw material. Working with this raw material rationally and systematically imposes the absolute prerequisite that the specialty elaborate, experiment, and explore the microbiology documentary language. This crucial job can be done only by an international community of microbiologists.

LITERATURE CITED

1. BRYGOO, P. R. 1963. Report of the Permanent Committee for Microbiological and Immunological Documentation, p. 705, 715. *In* N. E. Gibbons [ed.], Recent progress in microbiology. Symp. VIII Intern. Congr. Microbiol., Montreal, 1962. Univ. Toronto Press, Toronto.
2. BRYGOO, P. R., AND G. TUNEVALL. 1964. Documentation, classification and retrieval of information in applied microbiology, p. 102–112. *In* M. P. Starr [ed.], Global impacts of applied microbiology. Almqvist & Wiksell, Stockholm.
3. CLEVERDON, C. W. 1960. ASLIB Crainfield Research Project on the Comparative Efficiency of Index System. ASLIB Proc. **12**(12): 421–431.
4. CLEVERDON, C. W., AND J. MILLS. 1963. The testing of index language devices. ASLIB Proc. **15**(4):106–130.
5. FAIRTHORNE, R. A. 1958. Automatic retrieval of recorded information. Computer J. **1**:36–42.
6. GARFIELD, E. 1954. Citation indexes for science. Science **122**:108–111.
7. RAETTIG, H. 1958. Bacteriophagie. I, II. Gustav Fischer, Stuttgart.
8. RAETTIG, H. 1963. Poliomyelitis Immunität. I, II. Gustav Fischer, Stuttgart.